ROUTLEDGE-WIAS INTERDISCIPLINARY STUDIES

Corpus Methodologies Explained

An empirical approach to translation studies

Edited by
Meng Ji, Michael Oakes, Li Defeng and Lidun Hareide

Corpus Methodologies Explained

This book introduces the latest advances in Corpus-Based Translation Studies (CBTS), a thriving subfield of Translation Studies which forms an important part of both translator training and empirical translation research. Largely empirical and exploratory, a distinctive feature of CBTS is the development and exploration of quantitative linguistic data in search of useful patterns of variation and change in translation. With the introduction of textual statistics to Translation Studies, CBTS has geared towards a new research direction that is more systematic in the identification of translation patterns; and more explanatory of any linguistic variations identified in translations. The book traces the advances from the advent of language corpora in translation studies to the new textual dimensions and the shift towards a probability-variation model. Such advances in CBTS have enabled in-depth analyses of translation by establishing useful links between a translation and the social and cultural context in which the translation is produced, circulated and consumed.

Meng Ji is Associate Professor/Reader at the Department of Chinese Studies at the University of Sydney.

Lidun Hareide is Assistant Professor at Møreforsking AS, Volda, Norway.

Defeng Li is Professor of Translational Studies at the University of Macau, China.

Michael Oakes is Reader in Computational Linguistics at the University of Wolverhampton, UK.

Routledge-WIAS Interdisciplinary Studies
Edited by Hideaki Miyajima and Shinko Taniguchi,
Waseda University, Japan

1. Corporate Crime in China
History and contemporary debates
Zhenjie Zhou

2. Why Policy Representation Matters
The consequences of ideological proximity between citizens and their governments
Willy Jou, Luigi Curini and Vincenzo Memoli

3. Electoral Survey Methodology
Insight from Japan on using computer assisted personal interviews
Edited by Masaru Kohno and Yoshitaka Nishizawa

4. Corpus Methodologies Explained
An empirical approach to translation studies
Meng Ji, Lidun Hareide, Defeng Li and Michael Oakes

5. Clans and Religion in Ancient Japan
The mythology of Mt. Miwa
Masanobu Suzuki

Corpus Methodologies Explained

An empirical approach to translation studies

Meng Ji, Lidun Hareide, Defeng Li and Michael Oakes

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

WIAS 早稲田大学高等研究所
Waseda Institute for Advanced Study

First published 2017
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloguing-in-Publication Data
A catalog record for this book has been requested

ISBN: 978-0-415-71699-4 (hbk)
ISBN: 978-1-315-69412-2 (ebk)

Typeset in Galliard
by Out of House Publishing

Contents

Acknowledgements

Empirical translation studies represents a rapidly growing field of cross-lingual and cross-cultural studies. An important feature of recent development in empirical translation studies is the use of statistical research methods in the exploration of translational features at linguistic and textual levels. Recurrent patterns identified and extracted from quantitative translations bring valuable and much-needed insights into effective translation strategies and techniques to inform the teaching of practical translation, the development of translation theories and the design of new translation technologies and software to support cross-cultural communication. This book represents the joint effort of advancing empirical translation studies among four translation scholars from Australia, Norway, China and the UK.

The conceptualisation of this project was discussed and finalised among the co-authors when the first author of the book, Meng Ji, was affiliated with the Waseda Institute of Advanced Studies (WIAS), Waseda University, Tokyo, in 2012. As the first translation scholar to be awarded the prestigious WIAS Research Fellowship, she benefited greatly from the world-class research environment provided by WIAS, which was multi-disciplinary, stimulating and truly rewarding. As the first title of translation studies in the Routledge-WIAS Interdisciplinary Studies series, the publication of the book on the tenth anniversary of the foundation of WIAS reflects the tradition and aspiration of the world-leading research institute, i.e. to pursue research excellence to advance better cross-cultural communication and understanding.

List of tables

List of figures

Introduction to *Corpus Methodologies Explained*

An empirical approach to translation studies

by Meng Ji, Lidun Hareide, Defeng Li and Michael Oakes

Amidst the growing body of empirical translation studies and corpus translation studies in particular (CTS), the current volume represents the latest research in key areas of CTS such as machine translation (Chapter 1, Michael Oakes), translation genre variation and shifting (Chapter 2, Meng Ji), translation stylistics (Chapter 3, Defeng Li) and translation universals, including testing of the Gravitational Pull Hypothesis (Chapters 4–5, Lidun Hareide). The structural organization of the book is balanced between theoretical discussion and illustrative case studies. It aims to provide a focused introduction to the research paradigms which prevail in current CTS, i.e. from the development of statistical machine translation to the exploration of recurrent translational patterns called translation universals. From Chapter 1 to Chapter 5, the levels of theoretical postulation increase, as the research methods used gradually move from essentially corpus-driven (Chapter 1 and 2), via corpus-assisted (Chapter 3) to typical corpus-based translation studies (Chapter 4 and 5).

The distinction between these three main research paradigms within the current CTS, which is evolving rapidly, is largely based on the purposes and aims of the use of empirical evidence in the study of corpora. Throughout the book, the frequency-based analysis of language corpora, monolingual or multilingual, plays an instrumental role in the corpus analysis of translation. In corpus-driven translation research as exemplified by Chapter 1 (on statistical machine translation), and from a different perspective by Chapter 2 (on genre studies), corpus analysis tends to focus on the statistical modeling of linguistic and textual patterns which lead to the development of new computational language models, conceptual dimensions and analytical instruments in translation studies.

Chapter 1 offers an overview of important research paradigms in machine translation, i.e. rule-based machine translation, example-based machine translation, translation memories and statistical machine translation. The significance of this chapter is that it uses case studies in multiple languages to illustrate the rationale behind competing language and translation models. The linguistic analysis is enhanced with detailed explanations of relevant

statistical procedures which allow readers to obtain an in-depth understanding of machine translation systems from Google Translate to popular computer-assisted translation (CAT) language resources like translation memories.

Chapter 2 presents a quantitative analysis of contrastive distributional patterns of part-of-speech categories in monolingual English and Chinese corpora, and corpora which contain Chinese translations of English source texts. The corpus study adopts an essentially corpus-driven approach to the analysis of the quantitative data extracted from large-scale language corpora. The statistical analysis constructs three distinctive genre classification models for English, Chinese and translational Chinese as represented by the three large-scale corpora under study. The analysis shows that English written genres have a clear focus on techniques involved in the delivery of textual information. By contrast, the genre system of original Chinese gives more emphasis to language style rather than the delivery of actual textual information. The focus on the quality and stylistic features of the language implies that the prioritization of the aesthetic value of writing exists widely in the modern Chinese genre system, which is a long-standing tradition in the Chinese language and cultural system.

The exploratory statistical analysis of translational Chinese genres reveals that the genre system of translational Chinese is more complex than that of the original languages, as three sets of criteria have emerged in the corpus analysis which underline the configuration of the translational Chinese genre system. These are (1) features related to the communicative function of translation, i.e. explicitation, simplification and interactivity; (2) source-text oriented textual and linguistic features; and (3) target-text oriented textual and linguistic features. Such corpus findings suggest that translation is a highly purposed and complex system. If we consider translational textual features like explicitation, simplification and interactivity as essentially target-audience oriented translation strategies and tactics, the corpus-driven analysis in Chapter 2 seems to suggest that the contemporary Chinese translational genre system is overwhelmingly oriented towards the target language and culture.

Chapter 3 offers an overview of translation stylistics, an important area of corpus translation research. It deploys descriptive analyses widely used in corpus-based translation studies such as the type-token ratio, standardized sentence length variation and normalized word frequency lists to explore contrastive stylistic profiles of different target versions of a source text (the case study used is from two early English translations of the Chinese literary classic *Dream of the Red Chamber* or *Hongloumeng*). The methodological framework of Chapter 3 is distinct from that of Chapter 2 in that the frequency-based analysis used in Chapter 3 is largely descriptive, whereas the quantitative methods used in Chapter 2 are more exploratory, aiming to construct new analytical instruments to make necessary preparations for further theoretical development. If we could consider the type of corpus translation research exemplified by Chapter 2 as essentially corpus-driven,

the focus of the analytical strategies of Chapter 3 is to detect differences between paired translations and the source text. An important observation made in Chapter 3 regards the further analysis of the corpus findings at a social and cultural level; in other words, how to interpret the stylistic differences identified between different translations within the larger target social and cultural background – a methodological concern which points to the strengths and limitations of many similar studies on translation stylistics.

Chapters 4 and 5 reflect the theoretical branch of translation studies, which focuses on general tendencies in translations. These chapters offer two corpus-based studies of universally existent tendencies in translation, i.e. translation universals, which represent the main focus of corpus-oriented descriptive translation research. The study tests the previously untested Gravitational Pull Hypothesis (Halverson 2003, 2007, 2009, 2010). Since the Gravitational Pull Hypothesis intends to reconcile two seemingly opposing translation tendencies, full testing of this hypothesis entails testing of the mutually exclusive Over-representation of Target-Language Specific Features Hypothesis (Baker 1993, 1996) and the Unique Items Hypothesis (Tirkkonen-Condit 2001, 2004). Consequently, all three hypotheses posited on the suggested translation universal "over- or under-representation of target-language specific features" in translation studies are tested. In order to test these hypotheses, two comparable parallel corpora having the same target language but different source languages are needed. The feature to be tested must be unique to the target language in one of the language pairs, but must have a grammatical counterpart in the source language in the other language pair.

As a typical corpus-based study, Chapter 4 presents the design of the study, outlines the three hypotheses, the language pairs and the corpora used, as well as the grammatical structure that is tested in the case studies. In addition, Chapter 4 presents the first case study where the mutually exclusive Unique Items and Over-representation of Target-Language Specific Features hypotheses are tested. The Spanish gerund is used as the test object. In order to establish empirically that the Spanish gerund in fact constitutes a unique item in relation to Norwegian, a comparative study of 20 per cent of all of the Spanish gerunds in each text of the Norwegian-Spanish Parallel Corpus and their Norwegian counterparts is conducted.

Chapter 5 builds on the results from Chapter 4 in order to test the Gravitational Pull Hypothesis on the language pairs English-Spanish and Norwegian-Spanish, using the same grammatical structure (the Spanish gerund). The work presented in Chapters 4 and 5 demonstrates that the Gravitational Pull Hypothesis can be empirically tested using corpus data, and that the five core predictions of this hypothesis received support. In addition, the Unique Items Hypothesis was not upheld in translations from Norwegian, both with regards to frequent and to prototypical gerunds, and this raises important questions as to when this latter hypothesis applies and when it does not, and whether it is needed at all.

Since its inception in the 1980s, CTS has been one of the fastest growing research and teaching areas in translation studies as an independent academic discipline. The development of CTS owes much to the growing sophistication and specificity of related research methodologies. The current volume highlights three key research paradigms or sets of analytical strategies widely used in CTS: corpus-driven (statistical machine translation; exploratory corpus statistics), corpus-assisted (translation stylistics and parallel corpus comparison) and corpus-based (translation universal features or general translation tendencies) approaches. As the case studies used in each chapter demonstrate, each approach has its strengths and limitations, which reflects the very nature of empirical translation research. The delimitation of these three sets of distinct yet related research schemes contributes to the further expansion of the field, which relies to a large extent on the development of a robust, integrative and innovative methodological system for empirical translation research.

References

Baker, Mona (1993). Corpus linguistics and translation studies: Implications and applications. In Gill Francis, Mona Baker and Elena Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, pp. 233–250.

Baker, Mona (1996). Corpus-based translation studies: The challenges that lie ahead. In Harold L. Somers (ed), *Terminology, LSP and Translation: Studies in Language Engineering*. Amsterdam: John Benjamins, pp. 175–186.

Halverson, Sandra (2003). The cognitive basis of translation universals. *Target* 15(2): 197–241.

Halverson, Sandra (2007). Investigating Gravitational Pull in translation: The case of the English progressive construction. In Riita Jääskeläinen, Tiina Puurtinen and Hilkka Stotesbury (eds), *Text, Processes, and Corpora: Research Inspired by Sonja Tirkkonen-Condit*. Savonlinna: Publications of the Savonlinna School of Translation Studies 5, pp. 175–196.

Halverson, Sandra (2009). Elements of doctoral training: The logic of the research process, research design and the evaluation of design quality. *The Interpreter and Translator Trainer* 3(1): 79–106.

Halverson, Sandra (2010). Cognitive translation studies: Developments in theory and method. In Gregory M. Shreve and Erik Angelone (eds), *Translation and Cognition*. Amsterdam: John Benjamins, pp. 349–369.

Tirkkonen-Condit, Sonja (2001). Unique items – over – or underrepresented in translated language? In *The Third International EST Congress*, Copenhagen, Denmark.

Tirkkonen-Condit, Sonja (2004). Unique items – over – or underrepresented in translated language? In Anna Mauranen and Pekka Kujamäki (eds), *Translation Universals: Do They Exist?* Amsterdam/Philadelphia: John Benjamins, pp. 177–184.

1 The need for corpora in machine translation

Michael P. Oakes

Abstract

In this chapter we show that corpora, particularly parallel bilingual corpora, are essential in the development of automatic machine translation (MT) systems, whether translation memories, example-based or statistical. Specific topics examined are the Europarl corpus, similarity measures for sentence matching, the Hofland sentence aligner, automatic generalisation of translation examples through paraphrasing and the discovery of templates, statistical methods of building bilingual dictionaries, the development of MT for less-resourced languages and the evaluation of MT systems.

1. Introduction

This chapter will show that corpora, particularly parallel bilingual corpora, are almost the *sine qua non* of automatic machine translation (MT). In section 2 we will examine the four main paradigms in automatic MT, namely rule-based MT (the least dependent on corpora), translation memories (not strictly speaking "true" MT, but widely used by professional translators), example-based MT and statistical MT. In section 3 Europarl is described, a multilingual corpus built from transcripts of sessions of the European Parliament, especially for developing MT systems. In section 4 we describe how translation memory (TM) and example-based MT systems depend on finding the most similar stored examples to the sentence we wish to translate. This requires "matching", or the determination of how similar two sentences are to each other. Section 5 covers sentence-level alignment, or discovering automatically which sentence(s) of one language in a parallel corpus match which sentence(s) of the other. As a case study, we will consider Hofland's aligner, designed originally for English and Norwegian. Since gathering enough parallel corpus data can be a problem, in section 6 we discuss the automatic generalisation of translation examples – how can we make a single stored example represent a whole set of sentences? The techniques described include paraphrasing and the discovery of templates. In section 7 we look

at statistical methods of building the bilingual dictionaries, with frequency information, that are widely used in automatic MT. In section 8 the topic is the development of MT for "minority" or less-resourced languages, using Cebuano and Mapudungun as case studies. Finally, in section 9, we will look at how MT systems are evaluated – to help us identify the "best" system, and to learn which improvements are possible.

2. Paradigms for machine translation

In this section we will consider the main broad methods which have been used for MT. The earliest systems were called rule-based systems, because they were heavily dependent on language-pair specific rules. At about the same time, three other paradigms were introduced. Two of these, translation memories and example-based MT, both stored large numbers of previous translations against which new translations could be compared. The difference between them was that human translators took the final decision as to which parts of the previous translations could be reused, while in example-based MT, the machine decides which fragments to reuse. Statistical MT uses purely numeric data, derived from corpora, about the probabilities of the translations of individual words (which may have more than one counterpart in the other language) and the fluency of translated text as a function of the word adjacencies in it. While traditional statistical MT systems used information about individual word correspondences, a more recent development is to consider phrase correspondences across languages.

2.1 Rule-based machine translation

The earliest MT systems, prior to the 1990s, were called rule-based systems, and were built using linguistic knowledge in what Somers (2009) calls a rationalist approach. At that time corpora were relatively rarely used in the development of MT systems, not really coming into their own until the advent of what Somers describes as the data-driven or empirical approaches which came to the fore in the 1990s. However, many people at that time were looking at how the use of "controlled languages", where the range of vocabulary and allowed grammatical structures was both limited and fixed, could improve the performance of rule-based systems. The idea was that controlled languages would contain relatively little ambiguity, and thus would be easier for MT systems to process. Various groups at this time did make use of corpora to define the range of vocabulary and grammar that an MT system should work with, and thus they had (and have) a role in developing controlled languages. The TAUM group in Montréal used the set of words and structures in a 70,000-word corpus to define a controlled language for MT, and the Eurotra MT Project used the Europarl corpus for a similar purpose (Somers, 2009). We will briefly take a look at an example of a rule-based system which is taken from Arnold et al. (1993:76–77).

```
Sam likes London

[S [NP Sam][VP [V likes] [NP London]]]

Analysis

[S $1:H:like, $2:SUBJ, $3:OBJ]

Transfer

[S $1:H:plaire, $2:OBJ, $3:SUBJ]

Synthesis

[S [NP Londres][VP [V plait] [PP [P á][NP Sam]]]]

Londres plait á Sam
```

Figure 1.1 Rule-based translation from English into French

This example of the rule-based MT approach employs the transfer approach of which there were many variants. Here, the input sentence "Sam likes London" is input to a shallow parser, which produces a parse tree. A deeper parse still is required in the "analysis" phase, since we need to determine the subject and object of the sentence, as the translation of the English verb "like" into the French verb "plaire" requires that the positions of the subject and object be switched. This switch is effected in the "transfer phase", where the English dependency structure is replaced by the corresponding French one. The synthesis phase turns the French dependency structure into a shallow parse (shown in treebank notation) of the target language sentence. From this, the output "Londres plait à Sam" is generated.

To ensure that input texts are authored in a consistent way, written rules such as those by Pym (1990) were produced, imposing such constraints as "keep sentences short", "omit redundant words" and "avoid strings of nouns". To control the grammar, rules were given such as "verb particles are often ambiguous", and verbs with prepositions, which are also often ambiguous, should be rewritten as simple verbs. For example, "turn on" should be rewritten as "start" (Somers, 2003). AECMA (1995) produced an English lexicon for a controlled language for aircraft maintenance, as shown in Figure 1.2. Only approved words should be used, but for unapproved words an example is suggested.

Closely related to the idea of a controlled language is that of a "sublanguage", which is a subset of a whole natural language but with its own lexicon and syntax. One example of a sublanguage is "legalese". The difference between the two is that a controlled language is artificially imposed, while the restrictions of a sublanguage occur naturally (Somers, 2003). Nyberg et al. (2003) refer to "machine-oriented" (p. 246) and "human-oriented"

```
Approved word: prevent (v)
Definition: to make sure that something does not occur
Example: attach the hoses to the fuselage to prevent their movement
Unapproved word: preventative (adj)
Approved alternative: prevent (v)
Unapproved example: This is a corrosion preventative measure
Approved rewrite: This prevents corrosion
```

Figure 1.2 Two entries from the AECMA lexicon

controlled languages (p. 249). The most successful rule-based MT systems restricted themselves to sublanguages, such as the TAUM (Traduction Automatique de l'Université de Montréal) METEO system which translated weather bulletins for radio transmission from English into French (Grimaila and Chandioux, 1992). A more recent example is that the Caterpillar company has a controlled language in the domain of earth-moving machines, and a controlled language exists for the MT of Japanese patent information. Today the Smart Corporation (www.smartny.com) still specialises in establishing controlled language MT systems (Hutchins, 2011). Corpora can help in many ways in setting up controlled languages. Word frequency lists can easily be generated from corpora, and similarly the frequencies with which syntax rules are called upon can be found by parsing corpora and recording each rule as it fires. Sublanguage lexicons can be created by statistical comparisons of sublanguage corpora and reference corpora of the parent language.

Aikawa et al. (2007) performed an empirical evaluation to show that a controlled language can improve the quality of MT output, using a method which can be used for all types of MT, not only rule-based. They created their own set of controlled language rules, such as "don't use slang or colloquial expressions", and "maximum sentence length 25 words". They produced one set of input texts which adhered to the controlled language rules, and a corresponding one which violated them. Microsoft's MSR-MT statistical MT system, trained on texts in the domain of Information Technology, was used to translate English inputs into Arabic, Chinese, Dutch and French. They showed that the quality of MT output is inversely related to the post-editing effort, as measured by edit distance, in order to correct it, and in this way demonstrated that the translations of the controlled language-compliant texts were much better than the others. BLEU (see section 9.2.1) and human subjective appraisals (on a scale of "1: unacceptable, 2: possibly acceptable, 3: acceptable, 4: perfect") were also used as evaluation criteria. Interestingly, they also showed *which* controlled language rules had most impact on improved performance. Among the controlled language rules with most effect across the four target languages was the requirement for formal style. For example, "finish" was preferred to "wrap up", and was correctly translated into French as "terminer" as opposed to "empaqueter". A second important rule was that spelling should be correct,

because misspelled words would be unrecognised by the system and reproduced unchanged in the output. Correct capitalisation made sure that "Word" (the Microsoft product) was translated as "Word", while "word" with a lower case initial was translated into French as "mot". Reasons for the improvements in MT produced by controlled languages are that they increase the density of terms found in the corpus, so there is more chance of terms being found there and "learnt", and bigger numbers produce more accurate statistics. Note that in these experiments the controlled language is being used to produce a corpus, while in other work corpora are used to derive the controlled language.

2.2 Machine translation approaches which depend on parallel corpora

Although we have seen at least two ways in which corpora can help the process of rule-based MT, the real value of parallel corpora is seen in the types of MT systems used today. In fact, we will refer to translation memories (TM), example-based MT and statistical MT as being "corpus-driven", as they could not function without parallel corpora. Even early versions of these systems used small handcrafted corpora, or built them from examples of real sentences translated by the users of those systems.

Elena Frick (2006) lists a number of advantages of corpus-driven approaches to MT over rule-based MT. The system building cost is much less for corpus-driven approaches, as it is no longer necessary to handcraft large numbers of rules for syntax, semantic restrictions, structural transfer, word selection, sentence generation, and so on, a task which can only be done by trained linguists. Instead, we only need a large parallel corpus consisting of original sentences and their translations. The large number of rules in a rule-based MT system means that the computational running costs are much higher than for corpus-driven approaches. While the rules in a rule-based MT system are "hard-wired" into the system, and thus form an inseparable component, corpus-based approaches, in common with other effective systems in artificial intelligence, keep the knowledge (the parallel corpus) separate from the system which makes inferences from the knowledge (derives the translation). This means it is easy to transport the system to new domains simply by replacing the corpus. Additionally, the set of rules for rule-based MT must be based on some linguistic theory, while the use of a parallel corpus is theory independent. Rule-based MT is based on exact matching, and is thus unable to translate when the input cannot be matched exactly by the rules. In contrast, corpus-driven MT systems can work with inputs that are merely similar to the stored examples, and can return a reliability factor showing the degree of similarity between them. Finally, corpus-driven systems can easily be improved by adding suitable additional examples to the collection. In contrast, it is difficult to update a rule-based system, since many of the rules are dependent on each other, so changes will involve whole sets of interdependent rules.

New: Read the license agreement carefully, then **fill in and return the software registration card at the bottom of the license agreement.**

Old: **Fill in and return the software registration card at the bottom of the license agreement.** [78% match]

Füllen Sie die Software-Registrations-Karte unter den Lizenzvertrag aus, und senden Sie diese zurück.

New: Leave enough space around the computer to perform tasks such as inserting diskettes and accessing your printer, **monitor** and other optional equipment.

GER: Bildschirm, Monitor [No match]

Figure 1.3 Fuzzy matching and terminology recognition in TRADOS Translator's Workbench II

2.3 Translation memories

Translation memories (TMs) are now the most widely used technology supporting the translation industry (Reinke, 2003). Original texts and their human translations are stored, and typically broken down into convenient units such as sentences. Over time huge collections of parallel sentences are built up, and these can be "recycled" by matching them either exactly or partially with respect to a source language sentence which is to be translated. The advantages of TM systems are that they increase translators' productivity and ensure that terminology is used consistently. The idea of using the computer to help reuse human translations first appeared in the 1960s with a system built for the European Coal and Steel Community (ECSC) – it was essentially a bilingual keyword in context (KWIC) tool, but there were plans to retrieve similar translations in their contexts. The main components of a TM system have remained the same since the 1990s: the translation memory itself, a terminology management system, exact and partial matching, and a parallel concordancer. The sample output for the TRADOS Translator's Workbench shown in Figure 1.3 shows the action of both a terminology database and a TM database.

In the top screen, the sentence to be translated is labelled "New". This is matched against all the stored sentences in the TM, and the most similar sentence is retrieved, being labelled "Old". In section 4 we will look at a number of matching algorithms, and here the algorithm is able to identify a partial match of 78% between the "Old" and "New" sentences. The previous German translation of the best-matching "Old" sentence is displayed to the translator, who decides which portions of it can be reused in the formation of a German translation of the original "New" sentence. In the lower screen, no

closely matching "Old" sentence can be found for the "New" input sentence. However, the English term "monitor" is found in the system's terminology database, so at least it is possible to display the suitable German translations of this term, "Bildschirm" and "Monitor", which the translator might well want to incorporate into her translation of the "New" sentence.

In order to create TMs from parallel corpora, the corpora must be aligned – in practice, automatically. However, automatic alignments are rarely perfectly accurate, and so should be checked by human translators prior to use (Macdonald, 2001). The original idea was that translators would develop their own stores of useful and frequently required translation pairs, but nowadays pre-existing bilingual corpora are almost always used. O'Brien (1998:119) found that "a TM is always more accurate when created by interactive translation as opposed to automatic alignment", but felt that automatic alignment was adequate to start things off. Many TM systems now include software for aligning parallel texts at the sentence level. A problem with using "off-the-shelf" parallel corpora is that they may contain repetitions, leading to multiple matches, but this can also be a good thing as it shows whether certain phrases are frequently used and consistently translated. This type of frequency information is valuable in statistical MT, described in section 7. There is also value in showing different translations of a source sentence in different contexts, as it would be difficult to imagine all of these in advance (Somers, 2009). TM users are recommended to clear out useless sentences from time to time, either "never used" ones or those leading to bad translations (Somers and Fernandez Diaz, 2004).

Since it is a waste of time to translate material that has been translated before or is at least very similar to that which has been translated before, TM programs can free translators from repetitive work and allow them to do more creative tasks. This is particularly true in repetitive but commercially important domains such as in the translation of periodically updated technical documentation, where each version may differ only slightly from the last (Macdonald, 2001).

2.4 Translation memories and machine translation

TMs are not fully automatic MT systems, because it is the translator rather than the computer who must decide which parts of the retrieved target language sentences are to be used. In this respect they differ from the example-based MT systems described in section 2.5, which are able to translate without human intervention. TM systems can be integrated with fully automatic MT systems. All sentences which do not produce an exact or high-scoring fuzzy match with the TM can be sent for processing by fully automatic MT, then returned, possibly with a probability score, for human post-editing. Commercial systems such as Across or SDL Trados Studio include interfaces to both rule-based MT and statistical MT systems. For this integration to be smooth, the MT system must be trained with a sufficient quantity of

company-specific bilingual training text (Reinke, 2003). Reinke concludes that "[t]he field of computational linguistics has long ignored the relevance of TM as the major language technology used in professional translation" (p. 45), the two approaches of TM and fully automatic MT being worked on by largely different communities of researchers.

2.5 Example-based machine translation

Example-based MT was first developed in Japan, as is referred to in the seminal paper by Nagao (1984). It is now one of the main avenues of MT research. Nagao identified the three main components of example-based MT: firstly, fragments of text to be translated must be matched against a database of real examples (retrieval); then we find the corresponding translation fragments (alignment); and finally recombine these to produce the translated text (recombination). Example-based MT thus has two important and difficult steps beyond the simple matching task which it shares with TM. Recombination, in common with rule-based MT, can use a traditional grammar as a template, or one derived from a parallel corpus such as Wu's Stochastic Inversion Transduction Grammars (Wu, 1997). A difficulty with recombination is "boundary friction" where "fragments taken from one context may not fit neatly into another slightly different context" (Somers, 2009:1182). In the example given by Somers (2009) in the English-French translation pair "The old man is dead" / "Le vieil homme est mort", we can't simply swap "femme" for "homme", as we need gender agreement, which would also require replacement of "vieil" with "vielle", and "mort" with "morte". There is also the problem of overlap, such as when we try to combine the fragments "the operation was interrupted because" and "because the file was hidden" (Somers, 2009).

Example-based MT is closely related to TM, the main difference being that in example-based MT it is the computer rather than the translator that decides what to do with a found example (Somers, 2009). As with TM, example-based MT makes use of a parallel corpus of previous translations (the "example base"), portions of which are retrieved if they match the input text sufficiently well. As for TM, early example-based MT systems used handcrafted examples, but now use parallel corpora. Sometimes the corpora in example-based MT are annotated with part of speech (POS) information and tree banks. Compared with TM, example-based MT requires much linguistic or statistical pre-processing, including tagging and parsing, in order to process and extract suitable examples. These two approaches lie at opposite ends of a spectrum in memory-based translation (McTait and Trujillo, 1999).

In TM, examples are usually stored as linear, unannotated text, while a wide range of formats have been used for example-based MT. Since example-based MT originated as a variant of rule-based MT, early example-based MT systems such as that of Watanabe (1992) stored the examples as aligned tree

structures, such as those shown below for the Japanese-English pair "kanojo was kami ga nagai" / "she TOPIC hair SUBJ is-long" or "she has long hair".

```
[verb = nagai [wa = kanojo, ga=kami]]
[verb = have[subj = she, obj=hair [mod = long]]
nagai → have, long; kami → hair; kanojo → she
```

The lexicon below the Japanese and English tree structures shows how the trees align. The word "nagai" corresponds to both "have" and "long", because if another word governs "nagai" then its English translation should be connected to the word "have". Tree structures are not used much now, due to problems of storage space, and the computational overhead of parsing them during the translation itself. Later systems tend to annotate the examples in a more shallow fashion, such as with stemming or POS tags (Somers and Fernandez Diaz, 2004).

2.6 Statistical machine translation

Statistical MT systems, which were originally developed by Brown et al. (1990) at IBM, analyse very large parallel corpora of existing translations, learn their statistical properties and then use these to translate new input. Thus, the availability of corpora is key to the process. The corpora first need to be aligned both at the sentence and word level, and in recent developments, possibly the phrase level as well. There are three main components to a statistical MT system, the first of which is the translation model, which stores the probability of each word in the source language corresponding to each individual word in the target language, taking into account the fact that a single word in one language does not always translate as a single word in the other language, and also that there are sometimes differences in word order. The translation model tries to encapsulate the fidelity of a translation. The second component is the target language model, which tries to capture the fluency of a translation. Do certain sequences of words normally occur together in the target language? The language model and translation models are learnt from monolingual and bilingual corpora, respectively. The third component is a decoder, which considers all possible translations of the source sentence given the translation model and the target language model. This gives many possibilities, so we need to find the most probable or "best" of these (Somers, 2011). Foster et al. (2003) found that the choice of training corpus has a strong effect on the measured performance of statistical MT. A small corpus of within-domain training text produces better output than a larger one in the "wrong" domain. They also found that mixing several training corpora can be beneficial. A major evaluation campaign for statistical MT is ACL WMT, where the systems are trained on the Europarl corpus.

2.6.1 The translation model

Somers (2009) describes the theoretical situation where we know that every source word should be translated by a single target word, which is always the same. In such a case the translation model would be very simple: the set of target words, which most probably correspond to a set of source words, could be found by a simple dictionary. In reality, of course, a word in one language is not always translated by the same word in another. A simple example is "the" in English, which is translated into French as "le" about two thirds of the time, and "la" about one third of the time. Such information is held in probabilistic dictionaries, and we will see how to build these using statistical methods in section 7. In fact, the situation in real life is more complicated still, as a single word in the source language is not always translated by a single word in the target language, such as the English word "implemented", which can be translated into French as "mise en application". A word in the source language which has no equivalent in the target language is said to have a fertility of 0; one which corresponds 1:1 with its translation is said to have a fertility of 1; and one which maps onto two words (like "not" mapping onto both "ne" and "pas") is said to have a fertility of 2. A given word in the source language does not always have the same fertility with respect to the target language, so for each word we must empirically find the probabilities of the different fertilities it can take.

Both these components require that the source language sentences and their translations are first aligned at the word level, which is normally done using the EM algorithm (Koehn, 2010:88–92). Brown et al. (1995) described six variants of their model, the first three of which have been the basis of much future work. These three models take into account the various levels of complexity a translation model might have, as discussed above. The first assumes that a word and its translation occupy the same positions in both the source and target language, the second finds the relative "distortion" likelihoods of (source word, target word) position pairs, and the third includes fertility probabilities. GIZA++, often used to produce word-level alignment, is a package for implementing the various IBM models, and is downloadable from Franz-Josef Och's website at www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

2.6.2 The target language model

The target language model stores the probabilities of word n-grams which might occur in that language, as estimated by an analysis of monolingual corpora. The shortest n-grams which would take sequence data into account would be 2-grams, but Foster et al. (2003) used a 3-gram model, and four-word sequences are also used. The idea is that a frequently occurring, highly probable sequence such as "provides a gentle introduction" is a "better" target language phrase than the less likely sequence "a gentle